Received 22 January 2019 ▌Accepted : 26 April 2019 ▌Published : 26 April 2019
DOI : 10.20870/oeno-one.2019.53.2.2371

Unravelling the complexities of wine : A big data approach to yeast assimilable nitrogen using InfraRed spectroscopy and chemometrics

Gabriella Petrovic[1], Jose-Luis Aleixandre-Tudo[1,2] and Astrid Buica[1]

[1]Department of Viticulture and Oenology, P/Bag X1 Matieland, Stellenbosch 7600, South Africa
[2]Institute for Grape and Wine Sciences, P/Bag X1 Matieland, Stellenbosch 7600, South Africa

Corresponding author : abuica@sun.ac.za

ABSTRACT

Aim : Assimilable Nitrogen (YAN) has been identified as one of the main drivers of wine quality, influencing the production of various aromas and ensuring a successful fermentation to dryness. Due to the number of factors affecting YAN concentration and composition, paired with the complexities of yeast metabolism, more data is required to enable a comprehensive understanding of this important component of the grape juice matrix. The use of high throughput and information-rich techniques such as InfraRed spectroscopy can lead to a fast generation of a large amount of data. In addition, there is a possibility to maximise the information output of the generated data when combined with various descriptive and exploratory statistical techniques.
Conclusion : Given the recent developments in the fields of analytical equipment and chemometrics, the review explores the possibility of a Big Data approach for the research of one of the most important and versatile grape juice parameters, namely YAN

KEYWORDS

Big data, chemometrics, grapes, yeast assimilable nitrogen (YAN), InfraRed (IR) spectroscopy, calibration, wine

INTRODUCTION

Due to the growing power of the consumer in modern times (Deloitte Insight Report, 2014), together with the increasing awareness of food quality, safety and authenticity (Danezis *et al.,* 2016), increasing pressure is being placed on the world-wide wine market to become more innovative to keep up with consumer demands (Fleet, 2008 ; Pretorius and Bauer, 2002). This is illustrated by the growing gap in supply and demand – where firstly, a global decrease in wine consumption and an increase in wine production (mainly in the New-World countries) can be observed and secondly, the shift in consumer preferences towards more premium wines (Bisson *et al.,* 2002 ; Pretorius and Bauer, 2002 ; Swiegers *et al.*, 2005).

The 'technological push' that is required to ensure the success of the global wine industry can take many forms, but essentially, is based on the interaction between four primary streams of knowledge and technology, namely : chemistry, biology, mechanical technologies and scientific instrumentation (Smith, 2007). These knowledge streams are spread over both phases of the winemaking process ; *i.e.* viticulture and oenology. As briefly outlined by Smith (2007), the chemistry streams include aspects such as the chemistry of the soil, the subsequent chemical reactions taking place in the vine, as well as the production and interaction of various chemical elements present in the fermenting must and the final wine. The biological aspect entails an in-depth investigation into the biotic features such as the interactions between the various species of yeasts, bacteria and fungi on the grape as well as during the fermentation and maturation processes. Furthermore, mechanical technologies refer primarily to the machinery built to prune, harvest, destem, crush and ferment grape juice to wine, while scientific instrumentation incorporates the technology required for the monitoring and control of the grape, fermenting must and wine during maturation.

Thus, at the forefront of innovation in the wine industry lies the requirement for the deeper understanding of the interaction of the chemical and biological constituents involved during the various stages of the winemaking process, from vine to wine. This can, in turn, be facilitated by the development of efficient, accurate and cost-effective monitoring instrumentation and protocols.

This literature review will therefore start by touching on the progression of wine research in the pursuit of quality wine production and the important role that yeast assimilable nitrogen (YAN) plays in this respect. However, due to the multitude of factors affecting the YAN concentration and composition and the subsequent non-linear and synergistic interactions of the products of nitrogen metabolism, it is evident that, in order to allow a holistic understanding of the factors contributing to wine quality, a 'Big Data' approach is required. Therefore, this review will proceed by detailing the concept of Big Data and what is required for this field of wine research to become part of the 'Big Data revolution'. As such, the current and prospective methods for YAN quantification are reviewed for their ability to facilitate a 'Big Data' approach to wine flavour and quality.

WINE : A CONUNDRUM

Wine originated as a spontaneous process whereby the natural consortium of yeast present on the surface of the grape resulted in the conversion of sugars (glucose and fructose) into ethanol and carbon dioxide (Fleet, 2008). However, the understanding of this basic principle by Pasteur led to the desire of man to improve upon and control this process to their advantage and thus, by 1890, grape juice was being inoculated with pure yeast cultures (Fleet, 2008 ; Pretorius, 2000). This yeast, *Saccharomyces cerevisiae*, was selected based on its improved fermentative capacity and, subsequently, the possibility of a more predictable outcome (Fleet, 2008 ; Swiegers *et al.*, 2005).

As a result, a wealth of research has gone into understanding yeast and the conditions that are most conducive to the formation of a dry wine, free from spoilage. However, as this research field developed, the focus shifted towards making wines that exhibit more favourable organoleptic qualities (Fleet, 2003 ; Polášková *et al.*, 2008). Due to the complexity and variability of wine and the subjectivity of human perception, extensive investigations into how consumers perceive the quality of wine have been conducted (Bisson *et al.*, 2002 ; Fleet, 2003). However, at the heart of this lies the perception of the organoleptic characteristics of the wine. Thus, flavour – defined as a multisensorial construct that incorporates the

sensations of the ortho- and retro-nasal olfactory systems – has been widely accepted as the primary proxy of wine quality (Charters and Pettigrew, 2007).

Therefore, added pressure has been placed on the wine market to produce wines that are sensorially pleasing (Swiegers *et al.*, 2005). However, the investigation of aroma in wine is not an easy task due to the varying origins and the subsequent synergistic, non-linear interactions of these sensorially active compounds (Bisson *et al.*, 2002 ; Lambrechts and Pretorius, 2000 ; Polášková *et al.*, 2008 ; Styger *et al.*, 2011). Wine aroma is an amalgamation of varietal aromas (from compounds originating from the grape berry), pre-fermentative aromas (due to extraction and conditioning of the grape must), fermentative aromas (produced through the metabolic activities of the yeast and bacteria) and post-fermentative aromas (that evolve during ageing of the wine due to various chemical reactions in wooden barrels or after bottling). However, fermentative aroma compounds have been found to be the most important contributors to aroma and, as a result, the choice of the yeast together with the fermentation conditions, are the dominant factors in determining the aroma and subsequently, the quality of the final wine (Lambrechts and Pretorius, 2000 ; Polášková *et al.*, 2008 ; Rapp and Versini, 1991b ; Styger *et al.*, 2011). Therefore, as the contents of the grape berry and the resulting juice provide the nutrients required for the growth and fermentative activity of the yeast (and bacteria), the factors influencing the composition of these compounds become increasingly important in the context of quality wine production (Swiegers *et al.*, 2005).

The factors influencing the grape composition were reviewed by Jackson and Lombard (1993). These include various aspects, with varying degrees of control, many of which cannot be controlled at all – such as the macro- and meso-climate that the grapevine experiences – to factors such as micro-climate, soil and water and competition – which can be controlled up to a point by various viticultural practices such as canopy management, irrigation and fertilization programmes and pest, weed and disease management, respectively. Another very important factor that is reported is the genetics of both the grapevine and the rootstock which are considered to determine how the vine will react to all these aforementioned factors. Thus, the grape juice composition becomes the result of a multitude of intricate interactions, analogous to the complexity of a neural network.

WHAT IS "BIG DATA" ?

The term 'Big Data' has become a part of modern-day vocabulary, most commonly used in the field of business to facilitate the understanding of consumers. Nevertheless, the so-called 'Big-Data revolution' is just as indispensable to scientific research, providing the possibility of more data-driven and informed decision-making and hypothesis generation (Lusher *et al.*, 2014). There is, however, a rising concern among experts in this field of the understanding of what 'Big Data' really is as it is said to pave the way for the 4th industrial revolution (Yin and Kaynak, 2015). A common misconception is that size of the dataset is the only requirement that permits the use of this term (Jagadish, 2015). Therefore, many publications detailing the technicalities of big data have been made available (Gandomi and Haider, 2015 ; Jagadish, 2015 ; Kitchin, 2014 ; Lusher *et al.*, 2014 ; Yin and Kaynak, 2015).

The first attempt and widely accepted definition of Big Data was made by an analyst, Doug Laney from the META group. This definition came to be known as the 3Vs of Big Data : volume, velocity and variety. Later, two additional terms were added by IBM to characterise Big Data, these included value and veracity (Lusher *et al.*, 2014 ; Yin and Kaynak, 2015).

Volume

Simply put, volume refers to the magnitude of the dataset (Kitchin, 2014). However, there is a lot of debate around what constitutes a high-volume dataset and is said to be highly dependent on the field. Thus, defining a dataset as Big Data solely on the size is widely contested (Boyd and Crawford, 2012 ; Jagadish, 2015). The commonly applied criteria : that Big Data is a dataset that is 'too large to be managed by traditional methods' may not be relevant in the case of chemistry (Lusher *et al.*, 2014). Datasets collected in a field such as analytical chemistry – which requires the intentional measurement and analysis of a specific variable/compound – are orders lower than datasets from social media, for example.

Velocity

Velocity is the rate at which the data is generated (Gandomi and Haider, 2015). Generally, Big Data is defined as data that is continuously being generated, enabled by technology such as smartphones and sensors (Kitchin, 2014). Velocity is a critical aspect as no dataset can amount to 'Big Data' by any definition or scale if a means of obtaining the data quickly and efficiently does not exist. As chemical analyses are often complicated, time-consuming and expensive, there is a requirement of developing more easy-to-use, rapid and cost-effective means of generating data. This will facilitate the movement of analytical chemistry (and the subsequent fields for which the analysis is being conducted) to successfully enter the Big Data revolution.

Variety

It is by this definition that Big Data will enable the understanding of complex systems (Lehning *et al.*, 2009), such as those leading to the complex and unique flavour and quality of a wine. Obtaining a variety of data on the viticulture side is made relatively easy by infrastructure such as satellite and aerial imaging, weather stations and radars, gauge stations, ground and aerial Light Detection and Ranging (LiDAR), temperature and moisture sensors, etc. provided of course, that these are correctly placed and efficiently maintained (Kitchin, 2014). In other words, systems and technology to monitor and capture the 'cause' *i.e.* the factors causing a chemical/biological change in the grape juice matrix, have already been developed. However, as eluded to above in terms of the velocity of data generation, there is a gap in the available tools to efficiently measure the 'effect' *i.e.* the chemical properties of the grape juice, fermenting must or resulting wine.

Veracity

Veracity, which refers to the reliability of the data, is a particularly major challenge of Big Data in chemistry (Gandomi and Haider, 2015; Lusher *et al.*, 2014). Most frequently, errors originate from sample preparation, human error, problems with equipment, calibration, reporting, calculation errors and method selection Analytical Methods Committee Technical Brief No. 56, 2013; Ellison and Hardcastle, 2012). Therefore, data generated from different laboratories or by different operators have the possibility of being either unreliable, or if protocols are slightly modified, or a different instrument was used, this data may not be directly comparable (Lusher *et al.*, 2014). Considering these challenges, a collaborative effort is required from the scientific community to ensure the production of high-quality data and, where ever possible, the standardisation of protocols. A movement towards this can be seen in the implementation of VAM (Valid Analytical Measurements) principles or becoming ISO (International Organization for Standardization) accredited, which encourages the regular participation of laboratories in proficiency testing (Analytical Methods Committee Technical Brief No. 56, 2013). Thus, the adoption of these principals by laboratories conducting analysis on grape juice, must and wine is crucial to ensure the success of Big Data in the field of viticulture and oenology. However, due to the high level of specialization that is often required for certain analyses and the logical restrictions, the standardisation of operational protocols is practically impossible.

Value

The value attribute of Big Data is two-fold: generally, Big Data is described as having 'low value density' *i.e.* the original form of the data has a low value in relation to its volume; however, when processed, this data can impose a great deal of value on a process or activity (Demchenko *et al.*, 2013; Gandomi and Haider, 2015). Furthermore, a paradigm shift lies between traditional data generation and analysis compared to Big Data: rather than the intentional gathering of data for a predetermined purpose, Big Data seeks to find value and gain insights from the data itself (Kitchin, 2014). Although this approach can lead to a high value impact by unveiling hidden patterns present in the data, there is a concern that, due to the magnitude of the dataset, spurious correlations can be made (Gandomi and Haider, 2015). Thus, uncorrelated variables can erroneously be found to be correlated to one another (Fan and Lv, 2008), leading to false information and, subsequently, misinformed decision-making. This is a concern that affects all fields that make use of big data and should be acknowledged by the analyst. Fan and Lv (2008) suggest that reduction in the dimensionality of the data may help to mitigate this issue.

By reviewing the definition of Big Data, it becomes apparent that there are many dimensions in addition to just the sheer magnitude of the dataset. Furthermore, how the 5Vs of Big Data are interpreted is specific to the field. In other words, what constitutes as 'Big Data' in the field of science may not constitute as 'Big Data' in the field of business, due to the different constraints and logistics of the respective fields. Thus, at the crux of the holistic understanding of the winemaking process by the integration of wine research into the 'Big Data revolution' lies the need for the development of high-throughput and accessible techniques for chemical analysis.

YAN : A PRIMARY DETERMINANT OF WINE QUALITY

One of the most important components affected by these abovementioned factors, as reviewed by Jackson and Lombard (Jackson and Lombard, 1993), is the yeast assimilable nitrogen (YAN) concentration and composition (Bell and Henschke, 2005). As the principal yeast used for fermentation, *Saccharomyces cerevisiae*, does not exhibit sufficient extracellular proteolytic activity, it is thus not able to make use of larger peptides or grape proteins as a source of nitrogen. Thus, YAN primarily refers to α-amino nitrogen and ammonium ions, as these sources of nitrogen are able to easily pass through the yeast cell membrane (Bell and Henschke, 2005 ; Beltran *et al.*, 2004 ; Cooper, 1982 ; Henschke and Jiranek, 1993).

The importance of having sufficient quantities of easily assimilable nitrogen during fermentation is two-fold. Firstly, as nitrogen is required for the growth of the yeast cell by providing the necessary precursors required for protein and nucleic acid synthesis (Gobbi *et al.*, 2013), the concentration of available nitrogen significantly impacts the kinetics of the fermentation process (Bely, Sablayrolles and Barre, 1990 ; Bisson, 1999 ; Henschke and Jiranek, 1993). Thus, nitrogen deficiency has been highlighted as the primary cause for stuck/sluggish fermentations (Bisson, 1999). Secondly, the majority of fermentative aromas are affected by the concentration and composition of available nitrogen (Bell and Henschke, 2005 ; Ugliano *et al.*, 2007). The most significant impact that YAN has on wine flavour and aroma is by providing substrates (*i.e.* branched chain and aromatic amino acids) for the Ehrlich pathway (Hazelwood *et al.*, 2008). This pathway results in the formation of higher alcohols and through subsequent reactions, various esters and volatile acids (Styger *et al.*, 2011). However, YAN has been observed to not only impact the formation of aroma compounds for which it provides direct precursors, but also in the formation of various other compounds contributing to wine flavour and aroma such as organic acids (Torrea *et al.*, 2011) and terpenes (Carrau *et al.*, 2005). In other words, YAN can be seen as central and a dominating factor in the flavour and subsequently, quality of the final wine (Ugliano *et al.*, 2007).

Thus, it is no surprise that the role of YAN in the fermentation of grape juice to wine has been an area of research that has received increasing attention in the past three decades. A bibliometric search using the terms "Yeast Assimilable Nitrogen" AND "wine" OR "grape juice" OR "grape must" as a 'topic' in the 'Web of Science database' resulted in 3113 of a total 3928 papers that were published since 1990, with more than 100 papers published annually since 2005. The increasing interest in this topic was most probably fuelled by seminal papers which, to a great extent, laid the foundations and established the importance for nitrogen research in fermentation (Bely *et al.*, 1990 ; Henschke and Jiranek, 1993 ; Rapp and Versini, 1991a). Specifically, the synthesis of information by Rapp and Versini (1991b), reinforced the pivotal role that YAN plays in the formation of favourable flavours and aromas and subsequently, the connection that this has on the perceived quality of the resulting wine. These ideas were echoed in a more recent review published by Ugliano *et al.*, 2007.

Taking into consideration the varying origins and the multitude of factors influencing wine flavour and aroma, it is becoming clear that wine is a multi-faceted research field on the frontier of an array of disciplines such as viticulture, microbial ecology, chemistry and more recently, sensory science, in the pursuit of the production of a quality product, capable of meeting consumer demands (Swiegers *et al.*, 2005). In light of this, the field of wine research necessitates a collective effort to integrate all of these various streams of data in an effective and meaningful manner *i.e.* the wine research field needs to implement a 'Big Data' approach to facilitate the deeper understanding of all the interacting factors that are at play.

The collection of a large number of samples for the purpose of understanding the nitrogen dynamics in the grapevine and the subsequent nitrogen composition of the grape juice matrix has previously been reported (Butzke, 1998 ; Hagen *et al.*, 2008 ; Huang and Ough, 1991 ; Kliewer, 1970 ; Nicolini *et al.*, 2004 ; Nisbet *et al.*, 2014 ; Petrovic *et al.*, 2019 ; Spayd and Andersen-Bagge, 1996 ; Stines *et al.*, 2000).

These investigations were carried out in the form of surveys and have either examined the nitrogen content in terms of total YAN, free amino nitrogen (FAN) and ammonia (Butzke, 1998 ; Hagen *et al.*, 2008 ; Nicolini *et al.*, 2004 ; Nisbet *et al.*, 2014 ; Petrovic *et al.*, 2019) or have taken a deeper look into the FAN content by assessing individual amino acid concentrations (Huang and Ough, 1991 ; Kliewer, 1970 ; Petrovic, 2018 ; Spayd and Andersen-Bagge, 1996 ; Stines *et al.*, 2000). The results were mostly presented in a descriptive format – presenting the state of the nitrogen content of different cultivars, vintages and geographical origins in terms of average, maximum, minimum and median values. Furthermore, the number of samples above or below a pre-determined level (of total YAN or FAN) were also reported.

The surveys on the amino acid content of grape juices also generally followed this descriptive format. However, additional investigations using the amino acid data included whether a correlation of certain amino acids (such as proline or arginine or the proline : arginine ratio), or total α-amino nitrogen or total free α-amino acids could be correlated to the amount of total soluble solids (TSS) present at harvest (Spayd and Andersen-Bagge, 1996). No correlation could, however, be found. Huang and Ough (1991) proposed that the ratio of proline : arginine can be correlated to a specific cultivar and can thus be used to discriminate between different cultivars, although this hypothesis has yet to be tested. Furthermore, Stines *et al.* (2000) studied the changes in free amino acid profiles over the course of berry ripening as well as the distribution of various amino acids between the pulp, skin and seeds at harvest. The study concluded that, due to the high arginine content of the skins, fermentation efficiency could be improved by keeping the juice in contact with skins during fermentation.

One of the major findings from the surveys was that a large percentage of samples from various cultivars and origins suffer from nitrogen deficiency (total YAN < 140 mg N/L according to Bely *et al.*, 1990) and are thus not capable of supporting adequate growth of yeast during fermentation. No correlation could, however, be found between FAN and ammonia concentrations and YAN was found to be too variable to be used as an indicator of ripeness. Other the other hand, Nisbet *et al.* (2014) had some success in building cultivar-specific models for the prediction of total YAN at harvest based on pre-harvest YAN levels.

Although these surveys provided value in terms of describing the nitrogen status, a gap still exists in the understanding of the dynamics and factors affecting/resulting in a particular YAN concentration and composition. Due to the number of compounds contributing to the YAN status, it is not surprising that the influence of the factors may be more complex to predict. Due to the highly variable and complex nature of YAN, a greater number of samples throughout the growing season may be required. However, this is only the first step. Combining a large sample set with high-throughput analytical methods and efficient statistical means of extracting information can lead to a better understanding of the evolution of this particular component of the grape juice matrix.

Methods currently available to measure YAN

Methods which are most commonly used for the measurement of this important component of grape juice include the formol titration, nitrogen by *o*-phthaldialdehyde (NOPA), enzymatic ammonia and high-performance liquid chromatography (HPLC) (Gump *et al.*, 2002). The formol titration is a method that was first developed in 1907 by Sörensen for determining the protein concentration of samples. This method entails the addition of neutralized formaldehyde, for the purpose of liberating protons, which are subsequently titrated by sodium hydroxide to an end point, usually to a pH of 8.0. As this method does not react with imino acids (*i.e.* secondary amino acids) such as proline and hydroxyproline, it is useful for the measurement of YAN as these amino acids are generally not assimilable by yeast under fermentative conditions (Bell and Henschke, 2005 ; Gump *et al.*, 2002).

A method such as NOPA paired with enzymatic ammonia may be preferred as it enables the

determination of not only the total amount of YAN, but the proportion of FAN to inorganic nitrogen. NOPA is able to provide a measurement of the FAN content of the must through the derivatization of α-amino acid groups with *o*-phthaldialdehyde (OPA). This results in the formation of an isoindole derivative which is quantified using a spectrometer at 335 nm (Gump *et al.*, 2002). As imino acids are not able to form the required isoindole derivative, these amino acids are also not quantified by this method. Ammonia can be spectrophotometrically quantified at 340 nm through the reaction between glutamate dehydrogenase enzyme and the ammonium ion (Dukes and Butzke, 1998).

For a more comprehensive look into the nitrogen composition of the grape juice matrix, HPLC can be used for the quantification of individual amino acids and ammonia. The use of this method for the measurement of the amino acid content of the grape juice matrix was first proposed by Dukes and Butzke (1998). However, as previously mentioned, OPA is not able to react with imino acids and therefore, two methods have been proposed (and are currently in use) for the quantification of these amino acids in grape juice and wine ; these include the use of an additional derivatization agent, FMOCl (9H-fluoren-9-ylmethyl chloroformate) (Martínez-Rodríguez *et al.*, 2002), or the conversion of these (secondary) amino acids into primary amines. However, the use of AccQ•Tag as a derivatization reagent allows for the simultaneous derivatization of both ammonia and primary and secondary amino acids and is frequently paired with ultra-performance liquid chromatography for high resolution, rapid analysis (Armenta *et al.*, 2010). A full review of this topic can be seen in Callejón *et al.* (2010).

However, these methods are not suitable for 'Big Data' collection. This is primarily due to the complicated protocols required for sample preparation, instrument control and data interpretation (Liu *et al.*, 2011). Therefore, these methods can be rather labour intensive and subsequently, time-consuming. Further disadvantages of these conventional methods include the destruction of the sample material as well as posing a threat to the environment due to use of hazardous chemicals/reagents. As a result, the generation of chemical data is slow and usually only performed with a clear purpose or question in mind. Thus, there is a need for methods that require minimal to no sample preparation or reagents in order to provide rapid and cost-effective analysis of important components of the grape juice matrix, such as the available nitrogen.

SPECTROSCOPY IN WINE RESEARCH

Spectroscopy : A method for high-velocity data generation

The infrared (IR) region, found between the visible and microwave region of the electromagnetic spectrum, was first discovered by Herschel in 1800 (Cozzolino, 2009). The potential application of IR energy in chemical analysis was, however, only realised in 1882 by Abney and Festing, who correlated the absorption of certain wavelengths of light in this region to the presence of certain organic compounds (Thomas, 1991). Thus, an inference of the chemical composition of a particular substance/matrix can be made due to the vibrations (*i.e.* bending, stretching, rocking, scissoring and wagging) of the chemical bonds present and subsequently, the wavelengths of light that are absorbed versus the light that is either transmitted or reflected. This vibration of various chemical bonds at certain frequencies of IR energy is determined by properties such as the mass of the atoms, the shape of the molecule, the strength of the bonds between constituent atoms and the periods of the associated vibrational coupling (Blanco and Villarroya, 2002 ; McClure, 2003 ; Osborne *et al.,*1993). As a result, the need for derivatization and possibly separation, can be eliminated and subsequently, a method which is both rapid and cost-effective is provided due to the minimal (or possibly no) requirement for sample preparation or reagents (Bauer *et al.*, 2008 ; Cozzolino, 2015 ; Dambergs *et al.*, 2015 ; Gishen *et al.*, 2010 ; Liu *et al.*, 2011 ; Nicolaï *et al.*, 2007 ; Shah *et al.*, 2010). This was a ground-breaking discovery, addressing all the drawbacks of conventional methods and consequently, providing a means of high-velocity data generation that is required for Big Data collection.

Furthermore, another aspect that makes spectroscopy an effective tool in the context of Big Data is the possibility of the investigation of the matrix in its entirety (Cozzolino *et al.*, 2009). This has several advantages above traditional methods. Firstly, together with chemometrics, the complex interactions between the various

components present in the matrix can be taken into account while traditional chemical analysis tends to oversimplify the system by eliminating any interferences in the matrix (Geladi, 2003; Gishen *et al.*, 2010). This 'multivariate' approach is especially useful for a highly dynamic and complex matrix such as grapes, must and wine (Cozzolino *et al.*, 2009). Secondly, more than one parameter can be analysed at a time, amplifying the amount of data that can be generated (Bauer *et al.*, 2008; Gishen *et al.*, 2010). Thirdly, due to the non-destructive nature of spectroscopy, in situ analysis of the chemical composition of the grape, must or wine is made possible, thereby enabling effective and continuous monitoring of the process (Gishen *et al.*, 2010). **NIR vs. MIR**

The near infrared (NIR) and mid-infrared (MIR) ranges correspond to the wavenumbers 13400-4000 cm^{-1} and 4000-400 cm^{-1}, respectively (Blanco and Villarroya, 2002; McClure, 2003). The spectra obtained in the MIR range are due to the fundamental vibrations related to the stretching, bending and rotations of chemical bonds present in the matrix. Furthermore, the MIR region can be divided into four regions corresponding to the following wavenumbers: 4000-2500 cm^{-1} (X-H stretch, where X denotes O, N, C atoms), 2500-2000 cm^{-1} (triple bond), 2000-1500 cm^{-1} (double bond) and 1500-400 cm^{-1} (fingerprint region) (Blanco and Villarroya, 2002; Cozzolino, 2015; Osborne *et al.*, 1993). The fingerprint region is of particular interest for analysts testing the composition of various biological materials (for example, for the presence of water, proteins, lipids, fatty acids, nucleic acids and polysaccharides), as it allows for the unambiguous identification of chemical bonds. This is primarily due to the sensitivity of the bending and skeletal vibrations to large wavenumber shifts (Li-Chan, 2010).

NIR spectra on the other hand, are due to the complex overtones and combination bands of these fundamental vibrations occurring in the MIR range and therefore, peaks in the MIR range are often much sharper, offering higher resolution than the peaks found in the NIR range (Cozzolino, 2015). Overtones occur due to anharmonic transitions between non-contiguous vibrational energy states, whereas combination bands arise from simultaneous changes in energy due to the interaction of two or more vibrational modes (Blanco and Villarroya, 2002; Osborne, 2000). The bonds most frequently observed in NIR are C-H, O-H, N-H and S-H. This is due to the light weight of the hydrogen atom resulting in large changes in the dipole moment and, subsequently, large deviations from normal harmonic behaviour (Blanco and Villarroya, 2002). Consequently, the NIR spectrum is characterised by highly overlapping bands and which is said to hamper its ability to accurately measure analytes making up less than 1 % of the total matrix (Cozzolino, 2015; McClure, 2003). On the other hand, overlapping spectra can lead to a reduction in the number of wavelengths that are required for the analysis of a particular compound – a potential advantage of NIR over MIR (Cozzolino, 2015) which is being facilitated by the development of increasingly advanced instruments, computers and chemometric techniques (Gishen *et al.*, 2010; Kramer, 1998).

Applications of IR spectroscopy in wine research

Spectroscopy was first applied to wine in 1976 in the work done by Kaffka and Norris (1976). Their work entailed the analysis in transmission mode of a small sample set of spiked red and white wines. The samples were spiked with various compounds such as ethanol, tartaric acid and fructose. Through trial and error a set of wavelengths were eventually identified that could be used to build calibrations for the quantification of the various analytes, using MLR analysis (Sun, 2009). However, in subsequent years, the primary use of spectroscopy in the wine industry, was for the analysis of ethanol. This has since become a standard method, used for routine analysis (Gishen *et al.*, 2010).

As spectroscopy instruments improved and the field of chemometrics developed, a range of other parameters of grapes (intact berries, berry homogenates and juice/must) and wine (dry, sweet, dessert and fortified) have been investigated. The progression of this field has been extensively reviewed (Gishen *et al.*, 2005; Gishen *et al.*, 2010; Cozzolino *et al.*, 2006; Bauer *et al.*, 2008; Cozzolino *et al.*, 2011; Dambergs *et al.*, 2015, Wang *et al.*, 2017; dos Santos *et al.*, 2017). A lot of emphasis has been placed on investigating the possibility of quantifying total soluble solids (TSS), total acidity (TA), pH, anthocyanins, total polyphenol content, compounds which are routinely used to determine the quality and ripeness of the berries before harvest. The rationale for the

development of these calibrations is said to stem primarily from the lack of objective methods available for determining optimum harvest dates. This is a concern as the composition of the grape at harvest is accepted to be a major contributor to the quality of the final wine. Moreover, in finished wines, other than the ethanol content, the ability of spectroscopy to provide accurate readings of pH, volatile acidity, malic, tartaric and citric acid, glycerol, reducing sugars (glucose and fructose) as well as sulphur dioxide, have also been investigated.

In addition to quantification of important parameters present in grapes, juice/must and wine, these reviews report how spectroscopy can be useful for qualitative analysis in wine research. The reports included an array of applications ranging from predicting wine quality scores (white and red wines) as assessed by wine experts, to the adulteration of wines from various geographical origins as well as classifying the health of grapes as well as discriminating between various yeast strains.

There has been much less work into the viability of using this technology as a means to quantify YAN. This is most likely due to the fact that YAN is comprised of a range of different compounds which produce a distinctly weaker signal than what can be observed for major wine compounds such as ethanol and sugar. Thus, the task of building accurate calibrations for the quantification of YAN, FAN and ammonia is a much more daunting one. This can be seen by the unsatisfactory results reported in literature thus far. The first report for the quantification of assimilable nitrogen was by Manley *et al.* (2001). This study investigated the ability of FT-IR spectroscopy to quantify the FAN component of YAN by collecting 97 must samples from 6 different varieties over the course of two vintages. However, due to the large errors in prediction (SEP = 272.1 mg /L), rather than quantification, the study used the FAN values to discriminate between samples using Soft Independent Modelling by Class Analogy (SIMCA). Furthermore, nearly a decade later, using ATR-MIR spectroscopy, Shah *et al.* (2010) attempted to quantify total YAN as well as its components, FAN and ammonia, separately. Although this study collected a larger number of samples (n=350), these samples were only collected over a single vintage from a single winery. As such, the chances that these samples may not be representative of the variation contained by the greater population are relatively high. Skoutelas *et al.* (2011) aimed to provide a proof of concept for the use of FT-MIR spectroscopy for the quantification of total YAN. The partial least squares (PLS) calibrations showed very low errors in prediction (SEP = 5.9 mg N/L) and a very high RPD of 7.8. However, due to the lack of external validation and the removal of 40 % (n=28 of 71) of the samples (considered by the study to be outliers), the viability of FT-IR spectroscopy for the accurate quantification of total YAN was still inconclusive.

Given the success achieved for the calibration of a complex group of compounds such as phenolics (Aleixandre-Tudo *et al.*, 2015), which also have a markedly low signal in IR, as well as the central role that YAN plays in the production of quality wine, further research into this topic is warranted. However, careful consideration for the experimental design will be required. This will entail ensuring that a representative dataset is collected and that proper validation strategies are carried out to enable a realistic assessment of the predictive ability of the calibrations for the quantification of YAN, FAN and ammonia in grape juice.

Chemometrics and calibration

In order to extract value from infrared (IR) spectroscopy, a calibration needs to be set up. This can be achieved through multivariate data analysis techniques, also known as chemometrics, which facilitates the extraction of the analytical information contained by the spectra and correlates it to the properties contained by a set of reference data (Lavine and Workman, 2006; Wold, 1995). The calibration can either be qualitative, allowing the grouping of samples with similar characteristics *i.e.* the classification of unknown samples, or quantitative, where the concentration of a particular analyte can be predicted based on the on the spectral properties (Blanco and Villarroya, 2002).

However, before chemometric techniques are applied to spectroscopic data to predict the properties of new/unknown samples, there are a number of steps that need to be taken to ensure accurate predictions can be made. Therefore, before the various multivariate techniques are discussed, these steps and the rationale behind

them are briefly outlined in the following sections.

Gathering of calibration samples

The quality of the prediction is heavily dependent on the calibration set and, therefore, it is vitally important to ensure that the calibration set selected is representative of the population for which predictions are wished to be made (Blanco and Villarroya, 2002). The rationale for this stems from the fact that regression and classification methods used to build calibrations for spectroscopic instruments are essentially supervised learning techniques. In other words, the calibration set is the dataset that is used to train the model, *i.e.* the model learns from the information that is given to it in the form of the training set and, based on this, makes predictions on the properties of new samples that the model has not previously been exposed to (Wang *et al.*, 2012).

Due to the inherent variability of fruits and vegetables, building robust spectrophotometric calibrations for compositional analysis becomes a challenging task. Thus, the collection of a large number of samples from different 'batches' is crucial (Wang *et al.*, 1991). This means that careful consideration into what may cause variability in the sample needs to be taken into account to ensure that all this variability is well represented in the calibration set. For example, in the case of grapes for winemaking, variability may arise due to differences in cultivar and growing conditions and therefore geographical origin and vintage may also impact the variability of grape composition, in addition to the cultivar (Cozzolino, 2015 ; Nicolaï *et al.*, 2007 ; Sparrow *et al.*, 2015).

Dambergs *et al.* (2015) highlights the understanding of selecting an appropriate calibration (and validation) set as one of the largest barriers to the implementation of this technology into the wine industry. Thus, this is the first step for studies investigating the viability of this technology to provide accurate high-throughput compositional analysis for both wine research fields and the industry.

The use of an accurate reference method

Following the same rationale as above, whereby the calibration set is used to train the model and thus, the quality of the prediction is based on the quality of the calibration set, the method used to determine the reference concentrations (in the case of quantitative calibration) must be accurate (Blanco and Villarroya, 2002 ; S. Wang *et al.*, 2012). If reference methods are not carried out properly and produce values with large errors, it is possible that algorithms such as partial least squares (PLS) may still find correlations between these incorrect reference values and the spectra. This may lead to calibrations which seem accurate *i.e.* the reference data is faithfully represented by the model, however, in reality, the reference data does not faithfully represent the composition of the sample.

Recording of spectra

There are a range of considerations when deciding on which instrumentation to make use of, such as the properties of the sample to be analysed (solid/liquid/gas) as well as the appropriate wavelengths and resolution required for accurate analysis (Gishen *et al.*, 2010). The widespread application of IR spectroscopy is primarily due to the large degree of flexibility offered by these instruments, depending on the application, the characteristics of the samples, the conditions of the surrounding environment as well as the speed of data generation that is required (Blanco and Villarroya, 2002). Broadly speaking, an IR spectrophotometer consists of a radiation source (most commonly a tungsten halogen light bulb), accessories required for sample presentation, a monochromator, a detector, as well as a range of various optical components (optical fibres, beam splitters, integrating spheres and collimators) (Nicolaï *et al.*, 2007).

IR instruments can be grouped according to their wavelength selection properties *i.e.* whether they scan using the whole spectrum or only a limited set of fixed frequencies. Those with a limited set of frequencies either make use of filters or light emitting diodes (LEDs) and are generally simpler instruments, with limited resolution and no moving parts and are thus, generally used in portable instruments. Instruments employing the entire spectrum, generally referred to as scanning instruments, are more flexible and can therefore be used in a variety of applications. These scanning instruments can further be divided into monochromators, diode array and Fourier-transform (FT) spectrometers. In a scanning monochromator, the individual frequencies of light are separated by either a grating or a prism (Blanco and Villarroya, 2002 ; Gishen *et al.*,

2010 ; McClure, 2003 ; Nicolaï *et al.*, 2007). Photodiode array (PDA) spectrometers make use of a range of diodes emitting IR radiation and generally cover a range of 25000-5800 cm^{-1} (Osborne, 2000). There is widespread implementation of PDA spectrometers mainly due to the fast integration time and subsequent high acquisition speed, in addition to the absence of moving parts (Nicolaï *et al.*, 2007). Furthermore, FT spectrometers make use of an interferometer which modulates the radiation produced by the light source and is converted into a spectrum by means of a Fourier transform (Nicolaï *et al.*, 2007). There are two types of interferometers which are commonly used : a Michelson and a polarization interferometer, whereby the Michelson interferometer is said to produce the highest resolution (< 1 cm^{-1}) (Roberts *et al.*, 2004). Acousto-optically tunable filter (AOTF) is an additional type of monochromatic instrument, which makes use of an optical-band-pass filter that can be easily tuned to allow the passing of various wavelengths of radiation by adjusting the frequency of an acoustic wave moving through a crystal of TeO2 (Nicolaï *et al.*, 2007). Infrared spectrometers measuring in the mid-infrared range generally make use of an interferometer (FT) and attenuated total reflection (ATR) for sample presentation (Sorak *et al.*, 2012).

In addition to the types of radiation that the sample is exposed to, the extensive number of applications of IR spectroscopy in agriculture is owed to the range of different methods available for sample presentation (Osborne, 2000). In NIR spectroscopy, this includes transmittance, reflectance, as well as hybrids of the two phenomena, transflectance and interactance (Blanco and Villarroya, 2002 ; Nicolaï *et al.*, 2007 ; Osborne, 2000). For transmittance, the light source is placed opposite the detector. As radiation may either be absorbed, transmitted, or reflected by the sample of interest, when the intention is to collect spectra via transmittance, reflection is eliminated and therefore, the radiation attenuated by the sample may be interpreted as transmittance. The concentration of a particular analyte of interest can then be calculated via Beer-Lambert's law. However, this law becomes invalid in the case of light scattering, as the path length can no longer be defined due to the variation of light scattering from one sample to another. This is known as diffuse transmittance and is most commonly used for samples with a thickness of approximately 1-2 cm and is typically gathered in the range of 12500-9000 cm^{-1} (Nicolaï *et al.*, 2007 ; Osborne, 2000). In the case of reflectance, the radiation source and the detector are mounted at an angle to one another, such that the reflected radiation is recorded at an angle (for example, 45°). This is done to avoid specular reflection. Specular reflection is a phenomenon that occurs when all the radiation is reflected and therefore, no inference can be made about the chemical composition of the sample. Diffuse reflectance on the other hand, is when scattering causes the path length to be very large, resulting in an insignificant amount of transmittance and therefore, most of the incident light rays are reflected (Osborne, 2000). Transflectance is a modification of this phenomenon in the case of a liquid, where a ceramic tile is placed underneath the sample. As a result, the light is transmitted through the sample, reflected by the ceramic tile and transmitted back through the sample towards the detector. When the incident ray hits the sample surface and the resultant reflected ray is detected at a point adjacent to this incident ray, interactance takes place. This is achieved through the parallel placement of the light source and detector and is normally used for the analysis of large samples such as fruit (Osborne, 2000).

Attenuated total reflectance (ATR) is a technique that was developed by Fahrenfort (1961) to mitigate the issues associated with reflectance such as when substances show weak absorption but are also not suitable for transmission measurements. This was accomplished by using a dielectric with a high refractive index and the sample as the reflecting surface and as a result, the incident ray from the highly refractive dielectric (at an angle larger than the critical angle) will be 'totally' reflected. This will only occur at wavelengths where the sample is non-absorbing ; however, in the range where the sample is absorbing, there will no longer be total reflection, but instead, a highly contrasting and intense spectrum, similar to that of a transmission spectrum (Fahrenfort, 1961).

Furthermore, detectors in NIR spectroscopy can either be single or multiple channel. Single channel devices contain semiconductors of either PbS or InGaAs, whereas multiple channel devices contain a range of detection elements such as diode arrays (arranged in rows) or charged coupled devices (CCDs) (arranged in planes). These multi-channel devices are what

facilitate the simultaneous recording of a range of wavelengths and subsequently, responsible for the increased speed of spectra acquisition (Blanco and Villarroya, 2002).

As such, by taking all these options into account, it is clear that there are a range of factors that can affect the quality and stability of the response obtained by the spectrometer, necessitating the need for careful consideration when choosing an appropriate instrument for a specific application (Walsh *et al.*, 2000). **Pre-processing of spectra*

The aim of pre-processing is to remove any irrelevant information or physical phenomena that may hamper the subsequent classification, multivariate regression or exploratory data analysis techniques that may be applied to the data (Rinnan *et al.*, 2009 ; Roussel *et al.*, 2014). However, it should be kept in mind that pre-processing is not a solution for bad data collection, but rather for the inherent issues corresponding to a specific spectroscopic technique such as the base-line shifts and non-linearities strongly associated with IR spectra (Brown *et al.*, 2000 ; Rinnan *et al.*, 2009 ; Ruah *et al.*, 2014).

Broadly, the most popular pre-processing techniques can be classified into two groups : methods for scatter-correction and spectral derivatives (Rinnan *et al.*, 2009). Scatter-correction methods are used to lessen the spectral variability between samples induced by physical phenomena and include methods such as multiplicative scatter correction (MSC) and standard normal variate (SNV). Additionally, these methods have also been observed to correct for baseline shifts. In order to remove additive and multiplicative effects, spectral derivatives can be applied. When applying the first derivative, only the baseline is removed, whereas the second derivative also removes the linear trend in addition to the baseline. However, in practice, applying derivatives to raw spectral data generally results in noise inflation. To compensate for this, the Noris-Williams and Savitzky-Golay derivation techniques were developed which optimise the signal-to-noise ratio by smoothing of the spectra (Engel *et al.*, 2013 ; Nicolaï *et al.*, 2007 ; Rinnan *et al.*, 2009 ; Zeaiter *et al.*, 2005).

The most effective pre-processing technique is not easy to assess before model validation. However, Rinnan *et al.* (2009) give two pieces of advice in this regard : firstly, it is not advisable to apply too many pre-processing steps to a single data set and, secondly, essentially pre-processing should result in a reduction in model complexity. Furthermore, Engel *et al.* (2013) state that caution should be taken to avoid the introduction of additional variation in the data by pre-processing techniques. This statement stems from their investigation of a total of 4914 various pre-processing strategies, where only 5.6 % (273) were found to reduce model complexity and subsequently, increased the model accuracy. This result reiterates the importance of proper data collection to ensure accurate predictions, rather than relying on pre-processing.

Chemometrics

Without the development of chemometrics, IR spectroscopy would not have been as industrially relevant as it is today. Due to the inherent multivariate nature of IR spectra, statistical techniques considering more than one variable at a time needed to be developed (Bauer *et al.*, 2008b). Thus, in the late 1960s, extensive research was being done by an array of physical and analytical chemists to extract value from the multivariate responses obtained from these instruments and as a result, the field of chemometrics was born (Cozzolino *et al.*, 2009 ; Geladi, 2003 ; Wold, 1995). Consequently, chemometrics provides a means to examine as well as reveal important constituents through various interactions and interferences in the matrix (Geladi, 2003 ; Wold, 1995).

Chemometrics can be divided into two major categories : those used for quantitative analysis and those used for qualitative analysis (Blanco and Villarroya, 2002 ; Roussel *et al.*, 2014).

Quantitative methods

Quantitative analysis is mostly used for calibration purposes making use of regression techniques *i.e.* one/more dependent variables (Y-variables) are modelled based on a set of independent response variables (X-variables). Furthermore, regression analysis is essentially an example of supervised learning as ‘labelled’ training data (subsequently referred to as the calibration set) is used to make an inference about future ‘unlabelled’ samples (Olivieri, 2018). These methods are subsequently divided into linear and non-linear methods. The most frequently used methods include multiple linear

regression (MLR), principal component regression (PCR) and partial least squares (PLS) for linear methods and artificial neural networks (ANN) and non-linear PLS for non-linear methods (Blanco and Villarroya, 2002; Roussel *et al.*, 2014). However, this review will focus on briefly reviewing the linear methods.

MLR, developed by Norris in 1965, paved the way for quantitative chemometrics, however, it was not always successful at providing accurate predictions (McClure, 2003). This is mainly owed to its 'hard-modelling' approach which deals with the original variables and subsequently, assumes that the underlying chemical system is simple. In other words, the system is described in terms of a mathematical relationship whereby the measured variables are the independent variables and the outputs are the dependent variables. As a result, MLR is not robust against highly correlated (collinear), noisy data which may contain redundant (X variables). (McClure, 2003; Naes *et al.*, 2002; Wold *et al.*, 2001).

Due to these downfalls, soft-modelling approaches were designed by Wold (PLS) (1975) and Cowe and McNicol (PCR) (1985) which approach the regression problem from an entirely new angle. This approach assumes that the underlying chemical system is complex and therefore, soft-modelling (PLS and PCR) is based on the variation and correlation between the data points (*i.e.* the data found in the covariance matrix). Consequently, the interactions between variables as well as the overall variation in each of the independent variables can be taken into account (Geladi, 2003; Wold, 1995). The first step in this approach is to express the data as a set of latent variables *i.e.* the x-variables are projected onto a new set of axes which is based on the degree of variation that each x-variable contributes to and as a result, a new set of (uncorrelated) components are derived which are orthogonal to one another. The second step in the soft-modelling approach is to eliminate the components which do not explain an adequate amount of variation in the data *i.e.* an 'optimum' number of components needs to be selected. PLS is said to be superior to PCR in this regard, as the components selected in PCR are selected exclusively on the degree of predictor variance that is explained, whereas PLS seeks out the components that are most relevant in accurately predicting the outcome. This is important because if too many (unnecessary) components are selected, it may result in overfitting of the model and consequently, the model will not be able to accurately predict the properties/concentration of new samples as it is too reliant on the properties of the calibration/training set. This becomes especially relevant in small datasets where the number of components selected are more than the number of available samples. In light of this, the collection of a large number of samples which represents an adequate amount of variation present in the population becomes indispensable for accurate predictions of future samples (Geladi, 2003; Munck *et al.*, 1998; Naes *et al.*, 2002; Osborne, 2000; Reiss and Ogden, 2007; Wold, 1995).

In order to ensure that the regression model will result in accurate predictions of future samples, it is imperative to validate the model (Wold *et al.*, 2001). Methods currently used for method validation include internal (cross-validation) or external (test set) validation (Consonni *et al.*, 2010). Cross-validation can be defined as a validation technique that entails the division of the dataset into a predetermined number of subsets which are iteratively left out during calibration process, which is done until all the subsets have been left out once (Anderssen *et al.*, 2006; Hawkins *et al.*, 2003). Test set validation refers to the assessment of the predictive ability of the model by an independent set of samples which were not used to develop the calibration (Golbraikh and Tropsha, 2002).

Concerns have, however, been expressed among researchers in the field of chemometrics regarding the use of cross-validation as a measure of how accurately the model will predict future samples that the model has not yet "seen" (Anderssen *et al.*, 2006; Consonni *et al.*, 2010; Golbraikh and Tropsha, 2002; Gramatica, 2007). In a compelling study done by Golbraikh and Trophsa (2002), where several published datasets were investigated, it was shown that the R^2 obtained in cross-validation (often referred to as q^2) did not correlate with R^2-values obtained using an external test set. It was found that, often, the q^2-values were over-optimistic and when the datasets were tested with an external validation set, that the predictive ability was found to be considerably lower, yielding rather unsatisfactory results. Furthermore, Gramatica (2014) briefly overviews the arguments of

experts in the field (including his own), regarding best practices for model validation. Gramatica (2014) concludes that cross-validation and test set validation should not be viewed as alternatives but rather, used sequentially. The rationale for this is that cross-validation and test set validation have completely different aims: cross-validation should be used during model optimization to increase the robustness of the model and to preliminarily select the best models, whereas test set validation should be used for actual validation of the model (Consonni *et al.*, 2010; Gramatica, 2014). Ideally, the test set should become available to the modeller after the model has been developed, however, in practice, this is often not the case due to logistical issues and additional cost. Therefore, the best chance that the modeller has to verify the predictive ability of the available model is to exploit the data that is on hand *i.e.* randomly splitting the dataset into a test and calibration sets (Gramatica, 2014). This test set is therefore referred to as the external validation set as these samples will not at any time be exposed to the model during optimization, but rather be used to test the predictive ability of the model to predict future samples (Gramatica, 2014). However, the problem comes in with small datasets, where, if the dataset is split, that there is a chance that the dataset that is randomly selected is predicted well due only to chance (Consonni *et al.*, 2010; Hawkins *et al.*, 2003). In these cases, Hawkins *et al.* (2003), proposes that it is more statistically sound to do cross-validation; however, cross-validation procedures should be carried out wisely.

Nevertheless, when the appropriate validation technique has been selected based on the available data and considering the logistical constraints at play, there are a few model evaluation statistics which can be used to evaluate and report on the predictive ability of the regression model. The most popular is the squared correlation coefficient, R^2 (or q^2 in the case of cross-validation). This is owed to the easy comparison between models that this parameter offers, due to the independence of this value on the scale of the specific property that is being measured (in contrast to RMSEP or root mean square error in prediction, for example, which depends on the unit) (Consonni *et al.*, 2010). Instead, values universally range between 0 and 1 where 0 is indicative of the model not representing any of the variation present, whereas a value of 1 would indicate that the model accounts for the maximum amount of variation incorporated by the dataset (Consonni *et al.*, 2010). As such, more specifically, this value indicates how faithfully the variation that can be observed in the predictor variables (Y-variables) can be explained by the response variables (X-variables) in the calibration (R^2_{CAL} or q^2) and validation (R^2_{VAL} or R^2) sets (Aleixandre-Tudo *et al.*, 2018; Bauer *et al.*, 2008). Therefore, models with values closer to one are reported as having better predictive abilities and are therefore considered to be more accurate.

In addition to the squared correlation coefficient, the RMSEP (root mean square error in prediction) and RPD (residual predictive deviation) (RPD_{VAL}) can be calculated to evaluate the predictive ability of the model. This parameter is a measure of the mean deviation between the predicted and observed values (Consonni *et al.*, 2010). Thus, RMSEP is an estimate of the average uncertainty that is expected for the prediction of new samples not yet seen by the model (Nicolaï *et al.*, 2007).

The RPD is a ratio of the standard deviation incorporated by the dataset and the standard error of performance of the model and is therefore given by the following equation:\\Consequently, the more variability incorporated in the model (*i.e.* the higher the standard deviation) and the more faithfully the model is able to predict the outcome (*i.e.* the lower the RMSEP), the higher the RPD will be and therefore, the more reliable the model is thought to be. This is of course provided that the external validation set also incorporated enough variability to be representative of the population, allowing for realistic RMSEP values to be reported. In this case, a model with a high RPD will most likely be able to give a more accurate prediction of samples that it has not yet been exposed to. The rationale that a high standard deviation leads to more accurate prediction stems from the supervised approach of regression analysis where the model 'learns' from the characteristics presented to it in the training (calibration) set and therefore, if more information is used to train the model, the better it will be at making inferences/predictions of new samples.

Nicolaï *et al.* (2007), reviewed the RPD values that are relevant to PLS calibrations in

agricultural applications. RPD values between 1.5 and 2 are thought to be only sufficient to distinguish high values from low. Although RPDs between 2 and 2.5 allow for quantification, the level of quantification is considered only rough. For acceptable quantification purposes, values above 2.5 are required and values above 3 are preferable. Shah *et al.* (2010) regards RPD values ≥5 to be suitable for quality control for PLS calibrations for grape and wine analysis.

Qualitative methods

Other statistical methods that can be applied to chemical or spectral data are qualitative methods. These methods aim to classify an object (sample) rather than determining a quantitative property (Osborne, 2000). Fundamentally, these methods rely on developing a model based on pattern recognition strategies and can be divided into supervised and unsupervised techniques (Blanco and Villarroya, 2002).

Supervised methods can be divided into class-based models and discriminant analysis (DA) where class-modelling techniques focus on the similarities among samples in contrast to discriminant analysis which focuses on the differences (Blanco and Villarroya, 2002 ; Marini, 2010). The fundamental differences between these techniques are explained by Marini (2010) as follows : In the case of class-modelling, every class is modelled independently of the others ; accordingly, each sample is either accepted or rejected by the available classes. Consequently, when there is more than one class, a particular sample may only be accepted by one of the classes ; however, it is possible that the sample may be rejected by all the classes. In the case of overall rejection, this sample is identified as an outlier in terms of the available classes *i.e.* it may belong to a class that was not modelled. In contrast to this, discriminant techniques always assign a sample to one of the available classes. This is ensured by dividing the hyperspace of the available variables into as many segments as there are categories in the data. Therefore, if the coordinates of the sample fall into a particular segment which is labelled as "category 1" it will subsequently be assigned to that category. Examples of supervised methods for qualitative data analysis include Soft Independent Modelling by Class Analogy (SIMCA) supervised artificial neural networks (ANN), discriminant analysis (DA), partial-least squares discriminant analysis (PLS-DA) and its orthogonal version (OPLS-DA) and k-Nearest Neighbour (k-NN) analysis (Blanco and Villarroya, 2002 ; Roussel *et al.*, 2014 ; Siebert, 2011).

In terms of unsupervised methods, PCA has been acknowledged as one of the most indispensable chemometric techniques available (Cozzolino *et al.*, 2009 ; Siebert, 2011). The value in PCA stems from its ability to effectively screen, extract and compress multivariate data. This is achieved through a mathematical conversion of (potentially) correlated X-variables to a set of non-correlated variables which are orthogonal to one another. As a result, the dimensionality of the data can be reduced and the components explaining the maximum amount of variance present in the dataset can be identified. Therefore, based on whether samples group together or whether they separate from one another, hidden patterns in the data can be uncovered as well as allowing the detection of outliers (Cozzolino *et al.*, 2009 ; Naes *et al.*, 2002 ; Siebert, 2011).

Cluster analysis, another important unsupervised method for qualitative chemometric analysis, can broadly be divided into hierarchical, non-hierarchical and fuzzy clustering techniques (Siebert, 2011). The similarity between samples can be determined by various metrics including distances (Euclidean/Manhattan), correlations, as well as a combination of these. Most frequently, the samples are perceived as coordinates in a multidimensional space and the Euclidean distance between two samples are calculated ; the smaller the magnitude of the distance, the more similar the samples are considered to be. The fundamental difference between hierarchical and non-hierarchical clustering is whether a relationship among the clusters is established (hierarchical) or not (non-hierarchical). Therefore, in the case of hierarchical clustering the results are often represented as a dendrogram. Hierarchical clustering can further be divided into agglomerative (bottom-up) or divisive (top-bottom) approaches whereas non-hierarchical methods can be divided into partitioning, density-based, grid-based and 'other' (Gülağız and Şahin, 2017). Hierarchical and non-hierarchical clustering are, however, similar in terms of the assumption of single class-membership *i.e.* each sample may belong to only

one class. Conversely, fuzzy clustering algorithms allow samples to be members of two or more classes (Siebert, 2011).

The Soft Independent Modelling by Class Analogy (SIMCA) was the first supervised class-modelling method developed for the field of chemometrics (Marini, 2010). This method is in effect an extension of the unsupervised method, PCA and is often referred to as disjoint PCA (Bauer *et al.*, 2008). This is because the method groups objects together based on applying a PCA to each class of the training set. The ideal number of PCs can be determined by either double cross-validation or amount of explained variance or in some cases, it may be pre-determined (Rácz *et al.,*2018). Although SIMCA is a class-modelling technique, it is commonly used as a discriminatory tool in chemometrics. This is warned against by a meta-analysis conducted by Rácz *et al.* (2018), which shows that SIMCA was repeatedly outperformed for the task of discrimination by 29 different methods which includes the majority of the major categories of the available classification methods, such as linear and quadratic discriminant analysis (LDA), Classification and Regression Tree analysis (CART), PLS-DA, k-NN, to name a few.

CONCLUSION

Due to the multi-faceted nature of the winemaking process and the increasingly competitive world wine market, a need for more innovative technologies exists. These technologies will need to enable the accurate and continuous monitoring of various aspects of the process, from vine to wine. This is important as it will provide the tools and knowledge to increase the chances that a quality product can be produced.

Due to the highly complex and variable nature of YAN, 'traditional' wine research techniques appear to be lacking in providing a comprehensive understanding of the dynamics of this important component of the grape juice matrix. A 'Big Data' approach is thus suggested as a solution to the problem. However, in order to facilitate the integration of 'Big Data' in the field of wine research, methods for more rapid and cost-effective analyses are required. In light of this, IR spectroscopy, coupled with chemometrics, is recommended as a means to measure the YAN status of the grape juice matrix. This stems from the inherent features of speed, ease-of-use and lower costs associated with spectroscopy, in combination with the possibility of providing techniques for the multivariate assessment of complex systems, which is aided by chemometrics. Therefore, the field of chemometrics and spectroscopy could offer promising tools to facilitate the holistic understanding of complex systems, such as the nitrogen status of the grape juice matrix.

Acknowledgements : The authors would like to thank Winetech, NRF and THRIP for financial support.

REFERENCES

Aleixandre-Tudo J.L., Nieuwoudt, H., Aleixandre J.L. and du Toit W., 2018. Chemometric compositional analysis of phenolic compounds in fermenting samples and wines using different infrared spectroscopy techniques. Talanta, 176, 526–536. doi : 10.1016/j.talanta.2017.08.065

Aleixandre-Tudo J.L., Nieuwoudt, H., Aleixandre J.L. and Du Toit W.J., 2015. Robust ultraviolet-visible (UV-vis) partial least-squares (PLS) models for tannin quantification in red wine. Journal of Agricultural and Food Chemistry, 63(4), 1088–1098. doi : 10.1021/jf503412t

Analytical Methods Committee, AMCTB No 56., 2013. What causes most errors in chemical analysis ? Analytical Methods, 5(12), 2914–2915. doi : 10.1039/C3AY90035E

Anderssen E., Dyrstad K., Westad, F. and Martens H., 2006. Reducing over-optimism in variable selection by cross-model validation. Chemometrics and Intelligent Laboratory Systems, 84(1–2), 69–74. doi : 10.1016/j.chemolab.2006.04.021

Armenta J.M., Cortes D.F., Pisciotta J.M., Shuman J. L., Rasoloson D., Ogunbiyi O., … Shulaev V., 2010. A sensitive and rapid method for amino acid quantitation in malaria biological samples using AccQ•Tag UPLC-ESI-MS/MS with multiple reaction monitoring. Analytical Chemistry, 82(2), 548–558. doi : 10.1021/ac901790q.A

Barnett J.A., 2000. A history of research on yeasts 2 : Louis Pasteur and his contemporaries, 1850-1880. Yeast, 16(8), 755–771. doi : 10.1002/1097-0061(20000615)16:8<755::AID-YEA587>3.0.CO ; 2-4

Bauer R., Nieuwoudt H.H., Bauer F. F., Kossmann J., Koch K.R. and Esbensen K.H., 2008. FTIR spectroscopy for grape and wine analysis. Analytical Chemistry, 80(5), 1371–1379. doi : 10.1021/ac086051c

Bell S.-J. and Henschke P.A., 2005. Implications of nitrogen nutrition for grapes, fermentation and wine.

Australian Journal of Grape and Wine Research, 11(3), 242–295. doi : 10.1111/j.1755-0238.2005.tb00028.x

Beltran G., Novo, M., Rozès N., Mas A. and Guillamón J.M., 2004. Nitrogen catabolite repression in *Saccharomyces cerevisiae* during wine fermentations. FEMS Yeast Research, 4(6), 625–632. doi : 10.1016/j.femsyr.2003.12.004

Bely M., Sablayrolles J.M. and Barre P., 1990. Automatic detection of assimilable nitrogen deficiencies during alcoholic fermentation in oenological conditions. Journal of Fermentation and Bioengineering, 70(4), 246–252. doi : 10.1016/0922-338X(90)90057-4

Bely M., Sablayrolles J.M. and Barre P., 1990. Description of alcoholic fermentation kinetics : Its variability and significance. American Journal of Enology and Viticulture, 41(4), 319–324.

Bisson L., 1999. Stuck and Sluggish Fermentations. American Journal of Enology and Viticulture, 50(1), 107-119.

Bisson L.F., Waterhouse A.L., Ebeler S.E., Walker, M.A. and Lapsley J.T., 2002. The present and future of the international wine industry. Nature, 418(6898), 696–699. doi : 10.1038/nature01018

Blanco M. and Villarroya I., 2002. NIR spectroscopy : a rapid-response analytical tool. TrAC Trends in Analytical Chemistry, 21(4), 240–250. doi : 10.1016/S0165-9936(02)00404-1

Boyd D. and Crawford K., 2012. Critical questions for big data : Provocations for a cultural, technological and scholarly phenomenon. Information Communication and Society, 15(5), 662–679. doi : 10.1080/1369118X.2012.678878

Brown C.D., Vega-Montoto L. and Wentzell P.D., 2000. Derivative Preprocessing and Optimal Corrections for Baseline Drift in Multivariate Calibration. Applied Spectroscopy, 54(7), 1055–1068. doi : 10.1366/0003702001950571

Butzke C.E., 1998. Survey of Yeast Assimilable Nitrogen Status in Musts from California, Oregon and Washington, 49(2), 5.

Callejón R.M., Troncoso A.M. and Morales M.L., 2010. Determination of amino acids in grape-derived products : A review. Talanta, 81(4), 1143–1152.

Carrau F.M., Medina K., Boido, E., Farina L., Gaggero C., Dellacassa, E., Versini, G. and Henschke P.A., 2005. De novo synthesis of monoterpenes by *Saccharomyces cerevisiae* wine yeasts. FEMS Microbiology Letters, 243(1), 107–115. doi : 10.1016/j.femsle.2004.11.050

Charters S. and Pettigrew S., 2007. The dimensions of wine quality. Food Quality and Preference, 18(7), 997–1007. doi : 10.1016/j.foodqual.2007.04.003

Consonni V., Ballabio D. and Todeschini R., 2010. Evaluation of model predictive ability by external validation techniques. Journal of Chemometrics, 24(3–4), 194–201. doi : 10.1002/cem.1290

Cooper T.G., 1982. Nitrogen Metabolism in *Saccharomyces cerevisiae*, 61. In Strathern J.N., Jones, E.W., Broach, J.R. (Eds). The Molecular Biology of the Yeast Saccharomyces : Metabolism and Gene Expression, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 39– 99. DOI : 10.1101/087969180.11B.39.

Cowe, I.A. and McNicol J.W., 1985. The Use of Principal Components in the Analysis of Near-Infrared Spectra. Applied Spectroscopy, 39(2), 257–266. doi : 10.1366/0003702854248944

Cozzolino D., 2009. Near Infrared Spectroscopy in Natural Products Analysis. Planta Medica, 75(07), 746–756. doi : 10.1055/s-0028-1112220

Cozzolino D., 2015. The role of visible and infrared spectroscopy combined with chemometrics to measure phenolic compounds in grape and wine samples. Molecules, 20(1), 726–737. doi : 10.3390/molecules20010726

Cozzolino D., Cynkar W.N.S.G, Dambergs R. and Smith P., 2009. A brief introduction to multivariate methods in grape and wine analysis. International Journal of Wine Research, 1. doi : 10.2147/IJWR.S4585

Cozzolino D., Cynkar W., Shah N. and Smith P., 2011. Technical solutions for analysis of grape juice, must and wine : The role of infrared spectroscopy and chemometrics. Analytical and Bioanalytical Chemistry, 401(5), 1479–1488. doi : 10.1007/s00216-011-4946-y

Cozzolino D., Dambergs R.G., Janik L., Cynkar W.U. and Gishen M., 2006. Analysis of Grapes and Wine by near Infrared Spectroscopy. Journal of Near Infrared Spectroscopy, 14(5), 279–289. doi : 10.1255/jnirs.679

Dambergs R., Gishen M. and Cozzolino D., 2015. A review of the state of the art, limitations and perspectives of infrared spectroscopy for the analysis of wine grapes, must and grapevine tissue. Applied Spectroscopy Reviews. doi : 10.1080/05704928.2014.966380

Danezis G.P., Tsagkaris A.S., Camin, F., Brusic V. and Georgiou C.A., 2016. Food authentication : Techniques, trends and emerging approaches. TrAC Trends in Analytical Chemistry, 85, 123–132. doi : 10.1016/j.trac.2016.02.026

Deloitte Insight Report, 2014. The Deloitte Consumer Review : The growing power of consumers (Consumer Review).

Demchenko Y., Grosso P., de Laat C. and Membrey P., 2013. Addressing big data issues in Scientific

Data Infrastructure, 48–55. doi : 10.1109/cts.2013.6567203

dos Santos C.A. T., Páscoa R.N.M.J. and Lopes J.A., 2017. A review on the application of vibrational spectroscopy in the wine industry : From soil to bottle. TrAC Trends in Analytical Chemistry, 88, 100–118. doi : 10.1016/j.trac.2016.12.012

Dukes B.C. and Butzke C.E., 1998. Rapid Determination of Primary Amino Acids in Grape Juice Using an o-Phthaldialdehyde/N-Acetyl-L-Cysteine Spectrophotometric Assay. American Journal of Enology and Viticulture, 49(2), 125–134.

Ellison S.L.R. and Hardcastle W.A., 2012. Causes of error in analytical chemistry : results of a web-based survey of proficiency testing participants. Accreditation and Quality Assurance, 17(4), 453–464. doi : 10.1007/s00769-012-0894-2

Engel J., Gerretzen J., Szymańska E., Jansen J.J., Downey G., Blanchet L. and Buydens L.M.C., 2013. Breaking with trends in pre-processing ? TrAC Trends in Analytical Chemistry, 50, 96–106. doi : 10.1016/j.trac.2013.04.015

Fahrenfort J., 1961. Attenuated total reflection. Spectrochimica Acta, 17, 698–709.

Fan J. and Lv J., 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 70(5), 849–911. doi : 10.1111/j.1467-9868.2008.00674.x

Fleet G., 2003. Yeast interactions and wine flavour. International Journal of Food Microbiology, 86(1–2), 11–22. doi : 10.1016/S0168-1605(03)00245-9

Fleet G.H., 2008. Wine yeasts for the future. FEMS Yeast Research, 8(7), 979–995. doi : 10.1111/j.1567-1364.2008.00427.x

Gandomi A. and Haider M., 2015. Beyond the hype : Big data concepts, methods and analytics. International Journal of Information Management, 35(2), 137–144. doi : 10.1016/j.ijinfomgt.2014.10.007

Geladi P., 2003. Chemometrics in spectroscopy. Part 1. Classical chemometrics. Spectrochimica Acta Part B : Atomic Spectroscopy, 58(5), 767–782. doi : 10.1016/S0584-8547(03)00037-5

Gishen M., Cozzolino, D. and Dambergs R., 2010. The Analysis of Grapes, Wine and Other Alcoholic Beverages by Infrared Spectroscopy. Handbook of Vibrational Spectroscopy. doi : DOI : 10.1002/9780470027325.s8960

Gishen M., Dambergs R.G. and Cozzolino D., 2005. Grape and wine analysis - enhancing the power of spectroscopy with chemometrics. Australian Journal Of Grape And Wine Research, 11(3), 296–305. doi : 10.1111/j.1755-0238.2005.tb00029.x

Gobbi M., Comitini F., D'Ignazi G. and Ciani M., 2013. Effects of nutrient supplementation on fermentation kinetics, H2S evolution and aroma profile in Verdicchio DOC wine production. European Food Research and Technology, 236(1), 145–154. doi : 10.1007/s00217-012-1870-0

Golbraikh A. and Tropsha A., 2002. Beware of q2! Journal of Molecular Graphics and Modelling, 20, 269–276.

Gramatica P., 2007. Principles of QSAR models validation : internal and external. QSAR and Combinatorial Science, 26(5), 694–701. doi : 10.1002/qsar.200610151

Gramatica P., 2014. External Evaluation of QSAR Models, in Addition to Cross-Validation : Verification of Predictive Capability on Totally New Chemicals. Molecular Informatics, 33(4), 311–314. doi : 10.1002/minf.201400030

Gülağız F.K. and Şahin S., 2017. Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms, 9(1), 9.

Gump B.H., Zoecklein B.W., Fugelsang K.C. and Whiton R.S., 2002. Comparison of Analytical Methods for Prediction of Prefermentation Nutritional Status of Grape Juice. Am. J. Enol. Vitic., 53(4), 325–329. doi :

Hagen K.M., Keller M. and Edwards C.G., 2008. Survey of biotin, pantothenic acid and assimilable nitrogen in winegrapes from the Pacific Northwest. American Journal of Enology and Viticulture, 59(4), 432–436.

Hawkins D.M., Basak S.C. and Mills D., 2003. Assessing Model Fit by Cross-Validation. Journal of Chemical Information and Computer Sciences, 43(2), 579–586. doi : 10.1021/ci025626i

Hazelwood L.A., Daran J.-M., Maris A.J.A. van, Pronk J.T. and Dickinson J.R., 2008. The Ehrlich Pathway for Fusel Alcohol Production : a Century of Research on *Saccharomyces cerevisiae* Metabolism. Appl. Environ. Microbiol., 74(8), 2259–2266. doi : 10.1128/AEM.02625-07

Henschke P. and Jiranek V., 1993. Yeasts – metabolism of nitrogen compounds. in Wine Microbiology and Biotechnology. Harwood Academic Publishers : Chur, Switzerland. 77–164.

Huang Z. and Ough C.S., 1991. Amino Acid Profiles of Commercial Grape Juices and Wines. American Journal of Enology and Viticulture, 42(3), 261–267.

Jackson D.I. and Lombard P.B., 1993. Environmental and Management Practices Affecting Grape Composition and Wine Quality - A Review. American Journal of Enology and Viticulture, 44(4), 409–430.

Jagadish H.V., 2015. Big Data and Science : Myths and Reality. Big Data Research, 2(2), 49–52. doi : 10.1016/j.bdr.2015.01.005

Kitchin R., 2014. Big Data, new epistemologies and paradigm shifts. Big Data and Society, 1(1), 205395171452848. doi : 10.1177/2053951714528481

Kliewer W.M., 1970. Free Amino Acids and Other Nitrogenous Fractions in Wine Grapes. Journal of Food Science, 35(1), 17–21. doi : 10.1111/j.1365-2621.1970.tb12358.x

Kramer R., 1998. Chemometric techniques for quantitative analysis. New York : Marcel Dekker, Inc.

Lambrechts M.G. and Pretorius I.S., 2000. Yeast and its Importance to Wine Aroma - A Review. South African Journal of Enology and Viticulture, 21(Special Issue), 97–129.

Lavine B. and Workman J., 2006. Chemometrics. Analytical Chemistry, 78(12), 4137–4145. doi : 10.1021/ac060717q

Lehning M., Dawes N., Bavay M., Parlange M., Nath S. and Zhao F., 2009. Instrumenting the earth : next generation sensor networks in environmental science. The Fourth Paradigm : Data-Intensive Scientific Discovery, 45–51.

Li-Chan E.C.Y., 2010. Introduction to Vibrational Spectroscopy in Food Science. In J. M. Chalmers and P. R. Griffiths (Eds.), Handbook of Vibrational Spectroscopy. Chichester, UK : John Wiley and Sons, Ltd.

Liu F., He Y., Wang L. and Sun G., 2011. Detection of Organic Acids and pH of Fruit Vinegars Using Near-Infrared Spectroscopy and Multivariate Calibration. Food and Bioprocess Technology, 4(8), 1331–1340. doi : 10.1007/s11947-009-0240-9

Lusher S.J., McGuire R., van Schaik R.C., Nicholson C.D. and de Vlieg J., 2014. Data-driven medicinal chemistry in the era of big data. Drug Discovery Today, 19(7), 859–868. doi : 10.1016/j.drudis.2013.12.004

Manley M., van Zyl A. and Wolf E.E.H., 2001. The Evaluation of the Applicability of Fourier Transform Near-Infrared (FT-NIR) Spectroscopy in the Measurement of Analytical Parameters in Must and Wine. South African Journal of Enology and Viticulture, 22(2). doi : 10.21548/22-2-2201

Marini F., 2010. Classification Methods in Chemometrics. Current Analytical Chemistry, 6(1), 72–79. doi : 10.2174/157341110790069592

Martínez-Rodríguez A.J., Carrascosa A.V, Martín-Álvarez P.J., Moreno-Arribas V. and Polo M.C., 2002. Influence of the yeast strain on the changes of the amino acids, peptides and proteins during sparkling wine production by the traditional method. Journal of Industrial Microbiology and Biotechnology, 29(6), 314–322. doi : 10.1038/sj.jim.7000323

McClure W.F., 2003. 204 Years of near Infrared Technology : 1800–2003. Journal of Near Infrared Spectroscopy, 11(6), 487–518. doi : 10.1255/jnirs.399

Munck L., Nørgaard L., Engelsen S.B., Bro R. and Andersson C.A., 1998. Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. Chemometrics and Intelligent Laboratory Systems, 44(1–2), 31–60. doi : 10.1016/S0169-7439(98)00074-4

Naes T., Isaksson T., Fearn T. and Davies T., 2002. A user Friendly guide to Multivariate Calibration and Classification. Chichester UK : NIR Publications.

Nicolaï B.M., Beullens K., Bobelyn E., Peirs A., Saeys W., Theron K.I. and Lammertyn J., 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy : A review. Postharvest Biology and Technology, 46(2), 99–118. doi : 10.1016/j.postharvbio.2007.06.024

Nicolini G., Larcher R. and Versini G., 2004. Status of yeast assimilable nitrogen in Italian grape musts and effects of variety, ripening and vintage. Vitis - Journal of Grapevine Research, 43(2), 89–96.

Nisbet M.A., Martinson T.E. and Mansfield A.K., 2014. Accumulation and Prediction of Yeast Assimilable Nitrogen in New York Winegrape Cultivars. American Journal of Enology and Viticulture, 65(3), 325–332. doi : 10.5344/ajev.2014.13130

Olivieri A.C., 2018. Introduction to Multivariate Calibration : A Practical Approach. Springer International Publishing.

Osborne B.G., 2000. Near-Infrared Spectroscopy in Food Analysis. In R. A. Meyers (Ed.), Encyclopedia of Analytical Chemistry. Chichester, UK : John Wiley and Sons, Ltd.

Osborne B.G., Fearn T. and Hindle P.H., 1993. Practical NIR spectroscopy with applications in food and beverage analysis. Practical NIR Spectroscopy with Applications in Food and Beverage Analysis.

Petrovic G., 2018. A survey of the YAN status of South African grape juices and exploration of multivariate data analysis techniques for spectrometric calibration and cultivar discrimination purposes. MSc thesis, Stellenbosch University.

Petrovic G., Kidd M. and Buica A., 2019. A statistical exploration of data to identify the role of cultivar and origin in the concentration and composition of yeast assimilable nitrogen. Food Chemistry, 276, 528–537. doi : 10.1016/j.foodchem.2018.10.063

Polášková P., Herszage J. and Ebeler S. E., 2008. Wine flavor : chemistry in a glass. Chemical Society Reviews, 37(11), 2478–2489. doi : 10.1039/b714455p

Pretorius I.S., 2000. Tailoring wine yeast for the new millennium : novel approaches to the ancient art of winemaking. Yeast, 16(8), 675–729. doi : 10.1002/1097-0061(20000615)16:8<675::AID-YEA585>3.0.CO ; 2-B

Pretorius I.S. and Bauer F.F., 2002. Meeting the consumer challenge through genetically customized wine-yeast strains. Trends in Biotechnology, 20(10), 426–432. doi : 10.1016/S0167-7799(02)02049-8

Rácz A., Gere A., Bajusz D. and Héberger K., 2018. Is soft independent modeling of class analogies a reasonable choice for supervised pattern recognition ? RSC Advances, 8(1), 10–21. doi : 10.1039/C7RA08901E

Rapp A. and Versini G., 1991a. Influence of Nitrogen Compounds in Grapes on Aroma Compounds in Wine. Wein-Wissens, 51. doi : 10.1016/S0167-4501(06)80257-8

Rapp A. and Versini G., 1991b. Influence of nitrogen compounds in grapes on aroma compounds of wines. In Developments in Food Science (Vol. 37, pp. 1659–1694). Elsevier.

Reiss P.T. and Ogden R. T., 2007. Functional Principal Component Regression and Functional Partial Least Squares. Journal of the American Statistical Association, 102(479).

Rinnan Å., Berg F. van den and Engelsen S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry, 28(10), 1201–1222. doi : 10.1016/j.trac.2009.07.007

Roberts C.A., Workman J., Reeves J. B., American Society of Agronomy, Crop Science Society of America and Soil Science Society of America (Eds.)., 2004. Near-infrared spectroscopy in agriculture. Madison, Wis : American Society of Agronomy : Crop Science Society of America : Soil Science Society of America.

Roussel S., Preys S., Chauchard, F. and Lallemand J., 2014. Multivariate Data Analysis (Chemometrics). In C. P. O'Donnell C. Fagan and P. J. Cullen (Eds.), Process Analytical Technology for the Food Industry (pp. 7–59). New York, NY : Springer New York.

Ruah M.E.N.M., Rasaruddin N.F., Fong S.S. and Jaafar M.Z., 2014. Data preprocessing methods of FT-NIR spectral data for the classification cooking oil (pp. 890–897). doi : 10.1063/1.4903688

Shah N., Cynkar W., Smith P. and Cozzolino D., 2010. Use of attenuated total reflectance midinfrared for rapid and real-time analysis of compositional parameters in commercial white grape juice. Journal of Agricultural and Food Chemistry, 58(6), 3279–3283. doi : 10.1021/jf100420z

Siebert K.J., 2011. Using Chemometrics To Classify Samples and Detect Misrepresentation. In S. E. Ebeler G. R. Takeoka and P. Winterhalter (Eds.), Progress in Authentication of Food and Wine (Vol. 1081, pp. 39–65). Washington, DC : American Chemical Society.

Skoutelas D., Ricardo-da-Silva J.M. and Laureano O., 2011. Validation and comparison of formol and FT-IR methods for assimilable nitrogen in vine grapes. South African Journal of Enology and Viticulture, 32(2), 262–266.

Smith K., 2007. Technological and economic dynamics of the world wine industry : an introduction. International Journal of Technology and Globalisation, 3(2/3), 127. doi : 10.1504/IJTG.2007.014329

Sorak D., Herberholz L., Iwascek S., Altinpinar S., Pfeifer F. and Siesler H.W., 2012. New Developments and Applications of Handheld Raman, Mid-Infrared and Near-Infrared Spectrometers. Applied Spectroscopy Reviews, 47(2), 83–115. doi : 10.1080/05704928.2011.625748

Sparrow A.M., Dambergs R.G., Bindon K.A., Smith P.A. and Close D.C., 2015. Interaction of

Grape Skin , Seed and Pulp Tissues on Tannin and Anthocyanin Extraction in Pinot noir Wines. American Journal of Enology and Viticulture, 1–27. doi : 10.5344/ajev.2015.15022

Spayd S.E. and Andersen-Bagge J., 1996. Free Amino Acid Composition of Grape Juice From 12 *Vitis vinifera* Cultivars in Washington. American Journal of Enology and Viticulture, 47(4), 389–402.

Stines A.P., Grubb J., Gockowiak H., Henschke P.A., HøJ P.B. and Heeswijck R., 2000. Proline and arginine accumulation in developing berries of *Vitis vinifera* L. in Australian vineyards : Influence of vine cultivar, berry maturity and tissue type. Australian Journal of Grape and Wine Research, 6(2), 150–158. doi : 10.1111/j.1755-0238.2000.tb00174.x

Styger G., Prior B. and Bauer F.F., 2011. Wine flavor and aroma. Journal of Industrial Microbiology and Biotechnology, 38(9), 1145–1159. doi : 10.1007/s10295-011-1018-4

Sun D.-W., 2009. Infrared Spectroscopy for Food Quality Analysis and Control. Academic Press.

Swiegers J.H., Bartowsky E.J., Henschke P.A. and Pretorius I.S., 2005. Yeast and bacterial modulation of wine aroma and flavour. Australian Journal of Grape and Wine Research, 11, 139–173.

Thomas N.C., 1991. The early history of spectroscopy. Journal of Chemical Education, 68(8), 631. doi : 10.1021/ed068p631

Torrea D., Varela C., Ugliano, M., Ancin-Azpilicueta C., Leigh Francis I. and Henschke P.A., 2011. Comparison of inorganic and organic nitrogen supplementation of grape juice - Effect on volatile composition and aroma profile of a Chardonnay wine fermented with *Saccharomyces cerevisiae* yeast. Food Chemistry, 127(3), 1072–1083. doi : 10.1016/j.foodchem.2011.01.092

Ugliano M., Henschke P.A., Herderich M. J. and Pretorius I.S., 2007. Nitrogen management is critical for wine flavour and style, 22(6), 8.

Walsh K.B., Guthrie J.A. and Burney J.W., 2000. Application of commercially available, low-cost, miniaturised NIR spectrometers to the assessmentof the sugar content of intact fruit, 27, 1175–1186.

Wang L., Sun, D.-W., Pu, H. and Cheng J.-H., 2017. Quality analysis, classification and authentication of liquid foods by near-infrared spectroscopy : A review of recent research developments. Critical Reviews in Food Science and Nutrition, 57(7), 1524–1538. doi : 10.1080/10408398.2015.1115954

Wang S., Wu D. and Liu K., 2012. Semi-supervised Machine Learning Algorithm in Near Infrared Spectral Calibration : A Case Study to Determine Cetane Number and Total Aromatics of Diesel Fuels (pp. 308–311). IEEE. doi : 10.1109 /ICICTA.2012.84

Wang Y., Veltkamp D.J. and Kowalski B.R., 1991. Multivariate instrument standardization. Analytical Chemistry, 63(23), 2750–2756. doi : 10.1021/ac00023a016

Wold H., 1975. Soft Modelling by Latent Variables : The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. Perspectives in Probability and Statistics : Papers in Honour of M. S. Bartlett., 117–142.

Wold S., 1995. Chemometrics ; what do we mean with it and what do we want from it ? Chemometrics and Intelligent Laboratory Systems, 30(1), 109–115. doi : 10.1016/0169-7439(95)00042-9

Wold S., Sjöström M. and Eriksson L., 2001. PLS-regression : a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. doi : 10.1016/S0169-7439(01)00155-1

Yin S. and Kaynak O., 2015. Big Data for Modern Industry : Challenges and Trends [Point of View]. Proceedings of the IEEE, 103(2), 143–146. doi : 10.1109/JPROC.2015.2388958

Zeaiter M., Roger J.-M. and Bellon-Maurel V., 2005. Robustness of models developed by multivariate calibration. Part II : The influence of pre-processing methods. TrAC Trends in Analytical Chemistry, 24(5), 437–445. doi : 10.1016/j.trac.2004.11.023

